If Else

```
load("cancer.rda")
common_cancers <- c("Breast", "Lung", "Colon", "Rectum", "Prosta
cancer_new <- subset(cancer, (Year == "5 Year"| Year == "20 Yea
Type %in%common_cancers)
```

```
happy_complete <- happy[complete.cases(happy[,c("happy", "marita
```

```
data_basketball <- Aus_athletes[which(Aus_athletes$sport == "bas
```

```
data_basketball_rowing <- Aus_athletes[which(Aus_athletes$sport
c("basketball", "rowing")), ]
```

```
# to select columns throughout the dataframe
select(interviews, village, no_membrs, months_lack_food)
# to do the same thing with subsetting
interviews[c("village","no_membrs","months_lack_food")]
# to select a series of connected columns
select(interviews, village:respondent_wall_type)
```

```
filter(interviews, village == "Chirodzo")
```

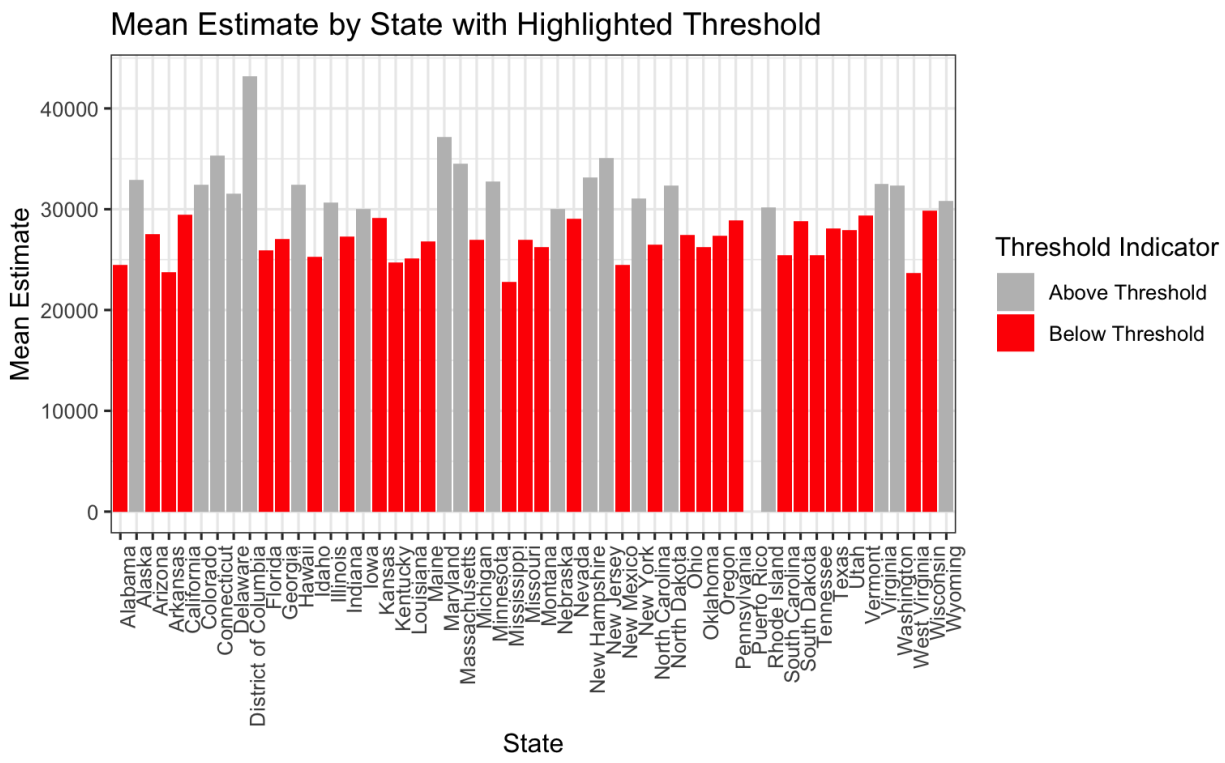
Frequently you'll want to create new columns based on the values in existing columns, for example to do unit conversions, or to find the ratio of values in two columns. For this we'll use `mutate()`.

We might be interested in the ratio of number of household members to rooms used for sleeping (i.e. avg number of people per room):

R

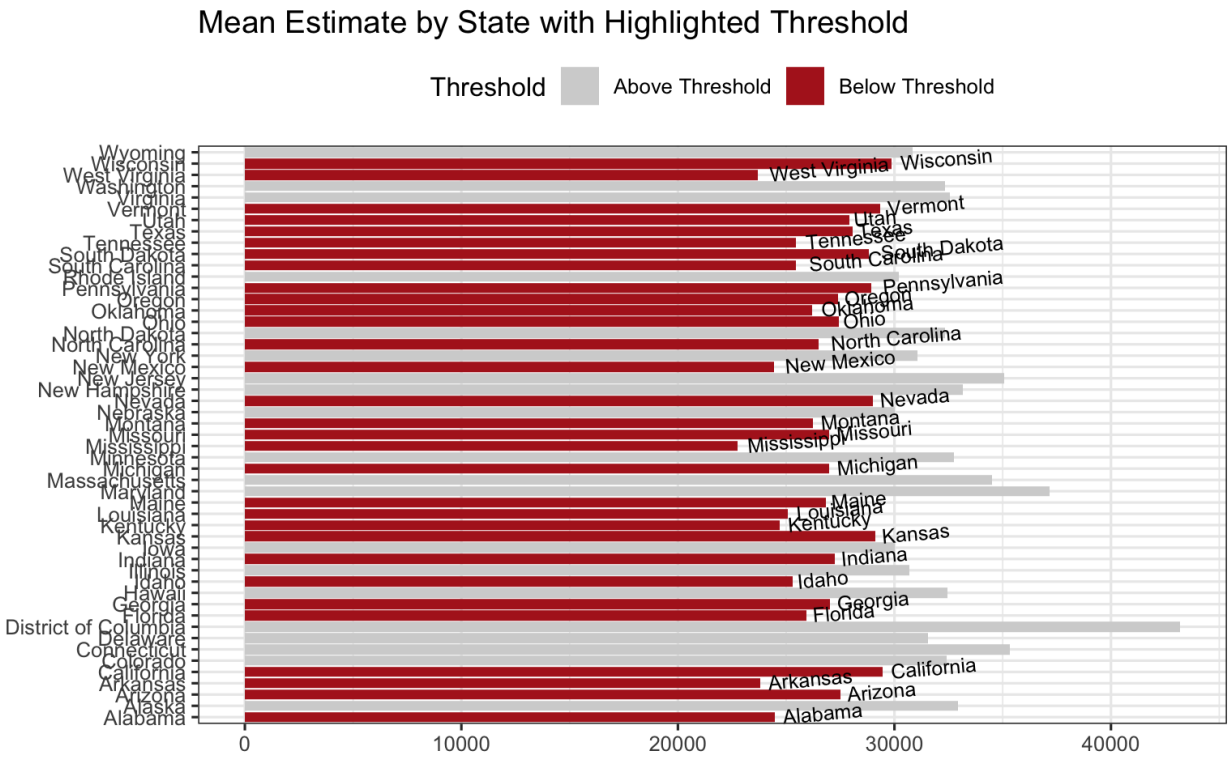
`interviews %>%  mutate(people_per_room = no_membrs / rooms)`

```
us_income <- us_rent_income[which(us_rent_income$variable == "in

threshold = 30000
us_income <- us_income %>%
  mutate(below_threshold = estimate < threshold)

# Plot the data
ggplot(data = us_income, aes(x = NAME, y = estimate, fill = belo
  stat_summary(fun = "mean", geom = "bar") +
  scale_fill_manual(values = c("TRUE" = "red", "FALSE" = "grey"
                    labels = c("TRUE" = "Below Threshold", "FALS
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(x = "State", y = "Mean Estimate", fill = "Threshold Indic
       title = "Mean Estimate by State with Highlighted Thresho
```




```r
library(ggplot2)
library(dplyr)

# Filter the data for the income variable
us_income <- us_rent_income %>%
  filter(variable == "income")

us_income <- na.omit(us_income)
# Define the threshold
threshold <- 30000

# Create a new column indicating whether the estimate is below t
us_income <- us_income %>%
  mutate(below_threshold = estimate < threshold)

# Calculate mean estimates per NAME
```

```
us_income_summary <- us_income %>%
  group_by(NAME) %>%
  summarize(mean_estimate = mean(estimate, na.rm = TRUE),
            below_threshold = any(mean_estimate < threshold))

# Plot the data
ggplot(data = us_income_summary, aes(x = NAME, y = mean_estimate
below_threshold)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = c("TRUE" = "firebrick", "FALSE" = "
                    labels = c("TRUE" = "Below Threshold", "FALS
                    "Above Threshold")) +
  geom_text(aes(label = ifelse(below_threshold, NAME, "")),
            hjust = -0.1, color = "black", size = 3, angle = 5)
  coord_flip() +
  labs(x = NULL, y = NULL, fill="Threshold", title = "Mean Estim
   by State with Highlighted Threshold") +
  theme(legend.position = "top") + theme(axis.title.x = element_
  axis.title.y = element_blank())
```


Mean Estimate by State with Highlighted Threshold
Threshold
Above Threshold
Below Threshold
Wyoming
Wisconsin
West Virginia
Washington
Virginia
Vermont
Utah
Texas
Tennessee
South Dakota
South Carolina
Rhode Island
Pennsylvania
Oregon
Oklahoma
Ohio
North Dakota
North Carolina
New York
New Mexico
New Jersey
New Hampshire
Nevada
Nebraska
Montana
Missouri
Mississippi
Minnesota
Michigan
Massachusetts
Maryland
Maine
Louisiana
Kentucky
Kansas
Iowa
Indiana
Illinois
Idaho
Hawaii
Georgia
Florida
District of Columbia
Delaware
Connecticut
Colorado
California
Arkansas
Arizona
Alaska
Alabama
0
10000
20000
30000
40000